# MICHAEL CULLAN

BROOKLYN, NY
(602) 301-0922
mjcullan@gmail.com

I was retained to review and analyze the expert disclosures of Mr. Maxwell Clarke. I am a statistician and machine learning engineer with published research in the practice of testing and comparing multiple explanations for historical data.

The analyses in question, described in Mr. Clarke's disclosures dated December 9, 2022 and December 22, 2022, apply statistical tests to a dataset of financials earnings announcements and trading events. Fisher's exact test, in particular, is poorly suited for the data at hand, which is collected from historical events rather than a carefully designed experiment. The simplistic assumptions of Fisher's exact test make it easier for the test to return an extreme result, even for data with unaddressed confounding variables, i.e. other relevant quantities which explain the patterns in the data, but which were excluded.

This issue is exacerbated by the choices made in collecting and filtering the data set. As described in Mr. Clarke's disclosure dated December 22, 2022,

> "Mr. Clarke then identified the number of earnings announcements that took place during the time that each of the eight traders was active.1 For example, between Vladislav Klyushin's first trade in this period (July 17, 2018) and his last (September 30, 2020), there were 38,359 earnings announcements."

This entire universe of earnings announcements is used to compute the results of the statistical tests. Every single earnings announcement is treated as equally relevant as a potential trading opportunity under the lens of the analysis applied. No justification is made that all possible data points should be used, as compared to a smaller subset of earnings events, which, for example, might have been restricted to companies in certain sectors or with particular stock performance patterns. These sorts of variables (e.g. sector, market capitalization) are not included in the dataset. As a result, they can not be used to evaluate alternative explanations. The decisions to include so many data points and to exclude other possibly relevant variables have strong implications on the analyses performed. In particular, they restrict the analyst to simplistic techniques which systematically produce more extreme results when conducted with larger data sets.

Finally, Mr. Clarke's analysis exaggerates the usefulness of its conclusions. In the following three statements, the disclosure dated December 22, 2022 presents an incorrect description of the interpretation of the obtained results:

1. "[Mr. Clarke] would opine based on a Fisher Exact test run on Mr. Klyushin's trading in this universe of earnings events that the probability that Mr. Klyushin would trade by chance almost exclusively in FA1 and FA2-serviced companies is at most 1 in a trillion, somewhat akin to flipping a coin 356 times and having the coin come up heads on 343 of the flips."

2. "Here, the likelihood that chance would lead to a situation where Mr. Klyushin's trading would correctly predict the most significant unexpected earnings outcomes was again less than 1 in a trillion.

3. "Mr. Clarke would opine that the observed relationship between the timing of their first trades in FA 2-serviced earnings events and the timing of the first download of information from FA 2 related those events would happen by chance (i.e., uncorrelated to the fact and timing of a download from FA2) less than 1 time in a million."

All three statements frame the results as the probability that the event in question would occur by chance, and the disclosure does not provide a name or definition of the types of quantities (e.g. "1 in a trillion") that are referenced. The quantities in question are called p-values. Mr. Clarke's expert disclosure interprets p-values in a manner inconsistent with their formal definition and recommended practice.

The American Statistical Association's *Statement on Statistical Significance and P-values* provides the following six principles:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

   *Reference:* Wasserstein, R. L., & Lazar, N. A. (2020). ASA statement on statistical significance and P-values. In The theory of statistics in psychology (pp. 1-10). Springer, Cham.

Mr. Clarke's expert disclosure repeatedly contradicts the second principle listed above, which states that p-values do not describe "the probability that the data were produced by random chance alone." By not referencing p-values by name, Mr. Clarke's disclosure further obscures this fact.

**Fisher's exact test**

Fisher's exact test, used in two of Mr. Clarke's analyses, was originally designed for simple, controlled experiments. In the case of these analyses, this test was applied to a large set of complex historical data. Each provided analysis simplifies the reality depicted in the data by comparing it against one variable at a time. The test outputs a statistically significant result, but it is asking a superficial question. For example, in the first analysis, it asks: if there were no non-random association between Mr. Klyushin's trades and the filing agencies, how likely would we be to see data this extreme? Answering this question is only relevant if the analyst has accounted for other explanations by including sufficient data.

Fisher's exact test was introduced by statistician Ronald Fisher in 1934 for the analysis of contingency tables. A contingency table summarizes a relationship between multiple categorical variables. For an example of how Fisher's exact test would be applied in an experiment, consider the following. Suppose I am growing 20 plants in identical conditions. I subject ten of them to freezing temperatures overnight. One week later, I see that five of the plants that had been left in the cold have died. I record my results in a table like so:

|  | Freezing | Non-freezing |
|---|---|---|
| Alive | 8 | 10 |
| Dead | 2 | 0 |

Empirically, a greater proportion of the freezing plants died, but a hypothesis test like Fisher's exact test helps to quantify the uncertainty in these observed results. For instance, if one applies Fisher's exact test to the table above, the p-value which, in this case, works out to 0.4737. This means that, assuming that the null hypothesis is true and that there is no non-random association between the freezing temperatures and the plants' dying, there would still be a 47.37% chance of observing such an extreme difference in death rates between the freezing and non-freezing groups.

On the other hand, suppose that five of the freezing plants had died instead of two. In that case, we would have obtained a p-value of 0.0325. If the freezing temperatures had no non-random association to the death rates, there would be a 3.25% chance that we would have observed such an extreme discrepancy.

In a controlled experiment, one can assure oneself that a reasonable number of other factors are accounted for. In this example, one can make sure all plants receive identical treatment except for the period of freezing temperatures. The experimental design supports the assumption that the table data adequately captures the conditions of the experiment.

It is much easier to account for other factors when analyzing data collected from designed experiments rather than historical observations. Continuing the plant example: what if I had collected these alive/dead from past observations, and in some cases, the plants might have been grown in different soil, or exposed to different amounts of sunlight, etc..? Applying Fisher's exact test on this historical plant data, only using the freezing/nonfreezing variable, means that the test's assumptions are less appropriate for the data. As such, the conclusions in this case would be less definitive.

**Interpreting p-values and statistical significance**

In a hypothesis test like Fisher's exact test, one starts by assuming, hypothetically, that there is no non-random relationship between the quantities of interest. This is called a null hypothesis. One then creates a statistical model to describe the data according to the assumptions of the null hypothesis. This null model is used to compute a test statistic, from which one ultimately obtains a p-value. The p-value describes the probability that, if the null hypothesis were true, we would see data as extreme as the data we are testing.

Because they are contingent on the null hypothesis, p-values do *not* measure the probability that the data were produced by random chance alone. Instead, they characterize the statistical compatibility of the data with the null hypothesis. This is a subtle difference in framing with major consequences in interpretation. If the test and null hypothesis are rooted in inappropriate assumptions or insufficient data, the results lose meaning even if the p-value concludes that they are statistically significant.

The classical approach to hypothesis tests is to start by specifying an allowable rate of false positives. As is stated in the expert's report, common choices for this rate are 1-in-20, or 0.05, and 1-in-100, or 0.01. If the p-value is smaller than the chosen threshold, the test is positive. So, for a test with a specified error rate of 0.05, a p-value of 0.0325 would give a positive result. This is what is typically meant by the term "statistical significance".  Hypothesis tests based on error rates provide a standard of admissibility for a statistical conclusion.

A p-value may be used to determine whether observed statistics are admissible, but this constitutes only one part of a sound statistical analysis. The overemphasis of p-values can be problematic, especially when dealing with large sample sizes. With a large sample, such as the one used in Mr. Clarke's analysis, even small discrepancies between groups can lead to extreme p-values.

Similarly, large sample sizes make it quite easy to obtain extremely low p-values even when grouping by arbitrary variables. As an example, I could cherry-pick an arbitrary variable from the earnings events data and use it to carry out a similar analysis to Mr. Clarke's first analysis. For example, each earnings event has an associated Central Index Key (CIK), which is a 10-digit ID number used on the SEC's computer systems to identify corporations and individual people who have filed disclosure with the SEC. If I group the events by whether or not the digits "32" appear in their CIK, I find that about 5.2% of events have "32" in their CIK, as compared to 1.1% of the events in which Mr. Klyushin traded. While this is an intentionally meaningless example with an apparently miniscule difference between the groups, it ends up being statistically significant. The p-value obtained via Fisher's exact test is less than 1 in 10,000.